Manipulating Attributes of Natural Scenes via Hallucination

LEVENT KARACAN, Hacettepe University and Iskenderun Technical University, Turkey
ZEYNEP AKATA, University of Tübingen, Germany
AYKUT ERDEM, Hacettepe University, Turkey
ERKUT ERDEM, Hacettepe University, Turkey

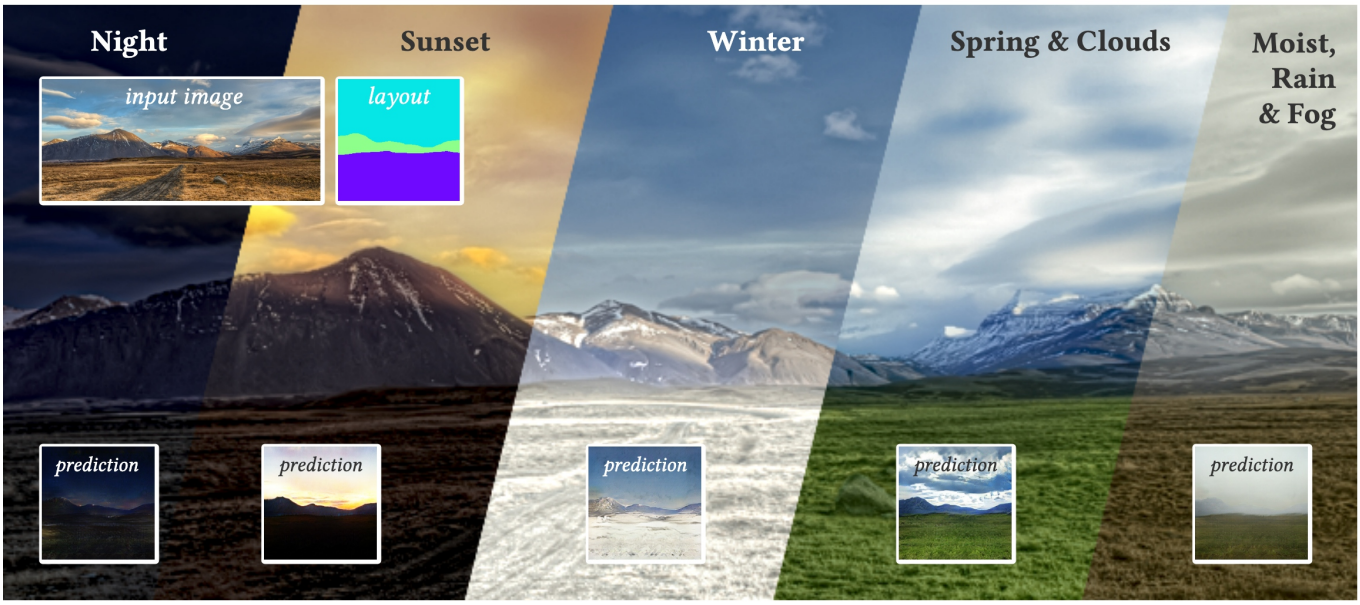


Fig. 1. Given a natural image, our approach can hallucinate different versions of the same scene in a wide range of conditions, e.g.*night*, *sunset*, *winter*, *spring*, *rain*, *fog* or even a combination of those. First, we utilize a generator network to imagine the scene with respect to its semantic layout and the desired set of attributes. Then, we directly transfer the scene characteristics from the hallucinated output to the input image, without the need for a reference style image.

In this study, we explore building a two-stage framework for enabling users to directly manipulate high-level attributes of a natural scene. The key to our approach is a deep generative network which can hallucinate images of a scene as if they were taken at a different season (e.g. during winter), weather condition (e.g. in a cloudy day) or time of the day (e.g. at sunset). Once the scene is hallucinated with the given attributes, the corresponding look is then transferred to the input image while preserving the semantic details intact, giving a photo-realistic manipulation result. As the proposed framework hallucinates what the scene will look like, it does not require any reference style image as commonly utilized in most of the appearance or style transfer approaches. Moreover, it allows to simultaneously manipulate a given scene according to a diverse set of transient attributes within a single model, eliminating the need of training multiple networks per each translation task. Our comprehensive set of qualitative and quantitative results demonstrate the effectiveness of our approach against the competing methods.

CCS Concepts: • **High-Level Image Editing** → **Image Processing**; • **Generative Adversarial Networks** → *Machine Learning*;

Additional Key Words and Phrases: Image generation, style transfer, generative models, visual attributes

ACM Reference Format:
Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. 2019. Manipulating Attributes of Natural Scenes via Hallucination. 1, 1 (October 2019), 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Authors' addresses: Levent Karacan, Hacettepe University and Iskenderun Technical University, Ankara, Turkey, karacan@cs.hacettepe.edu.tr; Zeynep Akata, University of Tübingen, Tubingen, Germany, zeynep.akata@uni-tuebingen.de; Aykut Erdem, Hacettepe University, Ankara, Turkey, aykut@cs.hacettepe.edu.tr; Erkut Erdem, Hacettepe University, Ankara, Turkey, erkut@cs.hacettepe.edu.tr.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2019 Association for Computing Machinery.
XXXX-XXXX/2019/10-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

> "The trees, being partly covered with snow, were outlined indistinctly against the grayish background formed by a cloudy sky, barely whitened by the moon."
> – Honore de Balzac (Sarrasine, 1831)

The visual world we live in constantly changes its appearance depending on time and seasons. For example, at sunset, the sun gets

close to the horizon gives the sky a pleasant red tint, with the advent of warm summer, the green tones on the grass leave its place in bright yellowish tones and autumn brings a variety of shades of brown and yellow to the trees. Such visual changes in the nature continues in various forms at almost any moment with the effect of time, weather and season. Such high-level changes are referred to as *transient scene attributes* – e.g. cloudy, foggy, night, sunset, winter, summer, to name a few [Laffont et al. 2014].

Recognizing transient attributes of an outdoor image and modifying its content to reflect any changes in these properties were studied in the past, however, current approaches have many constraints which limit their usability and effectiveness in attribute manipulation. In this paper, we present a framework that can hallucinate different versions of a natural scene given its semantic layout and its desired real valued transient attributes. Our model can generate many possible output images from scratch such as the ones in Fig. 1, which is made possible by learning from data the semantic meaning of each transient attribute and the corresponding local and global transformations.

Image generation is quite a challenging task since it needs to have realistic looking outputs. Visual attribute manipulation can be considered a bit harder as it aims at photorealism as well as results that are semantically consistent with the input image. For example, for predicting the look of a scene at sunset, visual appearances of the sky and the ground undergo changes differently, the sky gets different shades of red while the dominant color of the ground becomes much darker and texture details get lost. Unlike recent image synthesis methods [Chen and Koltun 2017; Isola et al. 2017; Qi et al. 2018; Wang et al. 2018], which explore producing realistic-looking images from semantic layouts, automatically manipulating visual attributes requires modifying the appearance of an input image while preserving object-specific semantic details intact. Some recent style transfer methods achieve this goal to a certain extent but they require a reference style image [Li et al. 2018; Luan et al. 2017].

A simple solution to obtain an automatic style transfer method is to retrieve reference style images with desired attributes from a well-prepared dataset with a rich set of attributes. However, this approach raises new issues that need to be solved such as retrieving images according to desired attributes and semantic layout in an effective way. To overcome these obstacles, we propose to combine neural image synthesis and style transfer approaches to perform visual attribute manipulation. For this purpose, we first devise a conditional image synthesis model that is capable of hallucinating desired attributes on synthetically generated scenes with semantic content similar to the input image and then we resort to a photo style transfer method to transfer the visual look of the hallucinated image to the original input image to produce a resulting image with the desired attributes.

A rich variety of generative models including Generative Adversarial Networks (GANs) [Goodfellow et al. 2014; Radford et al. 2016; Vondrick et al. 2016], Variational Autoencoders (VAEs) [Gregor et al. 2015; Kingma and Welling 2014], and autoregressive models [Mansimov et al. 2016; Oord et al. 2016] have been developed to synthesize visually plausible images. Images of higher resolutions, e.g. 256×256, 512×512 or 1024×1024, have also been rendered under improved versions of these frameworks [Berthelot et al. 2017; Chen and Koltun 2017; Gulrajani et al. 2016; Karras et al. 2018, 2019; Reed et al. 2016a,b; Shang et al. 2017; Zhu et al. 2017a]. However, generating diverse, photorealistic and well-controlled images of complex scenes has not yet been fully solved. For image synthesis, we propose a new conditional GAN based approach to generate a target image which has the same semantic layout with the input image but reflects the desired transient attributes. As shown in Fig. 1, our approach allows users to manipulate the look of an outdoor scene with respect to a set of transient attributes, owing to a learned manifold of natural images.

To build the aforementioned model, we argue the necessity of better control over the generator network in GAN. We address this issue by conditioning ample concrete information of scene contents to the default GAN framework, deriving our proposed attribute and semantic layout conditioned GAN model. Spatial layout information tells the network where to draw, resulting in clearly-defined object boundaries and transient scene attributes serve to edit visual properties of a given scene so that we can hallucinate desired attributes for input image in semantically similar generated image.

However, naively importing the side information is insufficient. For one, when training the discriminator to distinguish mismatched image-condition pairs, if the condition is randomly sampled, it can easily be too off in describing the image to provide meaningful error derivatives. To address this issue, we propose to selectively sample mismatched layouts for a given real image, inspired by the practice of hard negative mining [Wang and Gupta 2015]. For another, given the challenging nature of the scene generation problem, adversarial objective alone can struggle to discover a satisfying output distribution. Existing works in synthesizing complex images apply the technique of "feature matching", or perceptual loss [Chen and Koltun 2017; Dosovitskiy and Brox 2016]. Here, we also adopt perceptual loss to stabilize and improve adversarial training for more photographic generation but contrasting prior works, our approach employs the layout-invariant features pretrained on segmentation task to ensure consistent layouts between synthesized images and reference images. For photo style transfer, we use a recent deep learning based approach [Li et al. 2018] which transfers visual appearance between same semantic objects in real photos using semantic layout maps.

Our contributions are summarized as follows:

- We propose a new two-stage visual attribute manipulation framework for changing high-level attributes of a given outdoor image.
- We develop a conditional GAN variant for generating natural scenes faithful to given semantic layouts and transient attributes.
- We build up an outdoor scene dataset annotated with layout and transient attribute labels by combining and annotating images from Transient Attributes [Laffont et al. 2014] and ADE20K [Zhou et al. 2017].

Our code and models are publicly available at the project website[1].

[1] https://hucvl.github.io/attribute_hallucination

2 RELATED WORK

2.1 Image Synthesis

In the past few years, much progress has been made towards realistic image synthesis; in particularly, different flavors and improved versions of Generative Adversarial Networks (GANs) [Goodfellow et al. 2014] have achieved impressive results along this direction. Radford et al. [2016] were the first to propose a architecture that can be trained on large scale datasets, which sparked a wave of studies aimed at improving this line of work [Arjovsky et al. 2017; Mao et al. 2016; Salimans et al. 2016]. Larsen et al. [2016] integrates adversarial discriminator to VAE framework in an attempt to prevent mode collapsing. Its extension [Shang et al. 2017] further tackles this issue while improving generation quality and resolution. More recently, Karras et al. [2018] have suggested to use a cascaded set of generators to increase both the photorealism and the resolution of generated images. In the subsequent work, Karras et al. [2019] have achieved further improvement in realism and diversity of the generated synthetic images by adopting ideas from style transfer literature [Huang and Belongie 2017].

Conditional GANs (CGANs) [Mirza and Osindero 2014] that leverages side information have been widely adopted to generate images under predefined constraints. For example, the recently proposed BigGAN [Brock et al. 2019] generates high quality, high resolution images conditioned on visual classes in ImageNet. Reed et al. [2016a,b] generate images using natural language descriptions; Antipov et al. [2017] follow similar pipelines to edit a given facial appearance based on age. Pix2pix [Isola et al. 2017] undertakes a different approach to conditional generation that it directly translates one type of image information to another type through an encoder-decoder architecture coupled with adversarial loss; its extension Cycle-GAN [Zhu et al. 2017a] conducts similar translation under the assumption that well-aligned image pairs are not available. The design of our image synthesis model resembles CGANs, as opposed to Pix2pix, since those so-called image-to-image translation models are limited in terms of output diversity.

In the domain of scene generation, the aforementioned Pix2pix [Isola et al. 2017] and Cycle-GAN [Zhu et al. 2017a] both manage to translate realistic scene images from semantic layouts. However, these models are deterministic, in other words, they can only map one input image to one output image in different domains. Recently, some researchers have proposed multimodal (e.g. BicycleGAN [Zhu et al. 2017b]) or multi-domain (e.g. StarGAN [Choi et al. 2018], MUNIT [Huang et al. 2018]) image-to-image translation models. Both of these approaches have the ability to translate a given input image to multiple possible output images with the use of a single network. However, in BicycleGAN, the users have no control over the generation process other than deciding upon the source and target domains. StarGAN and MUNIT can perform many-to-many translations but these the translations are always carried out between two different modalities. Although these works improve the diversity to a certain degree, they are still limited in the sense that they do not allow to fully control the latent scene characteristics. For instance, these methods can not generate an image with a little bit of sunset and partly cloudy skies from an image taken on a clear day. Our proposed model, on the other hand, allows the users to play with all of the scene attributes with varying degrees of freedom at the same time.

Alternatively, some efforts on image-to-image translation has been made to increase the realism and resolution with multi-scale approaches [Chen and Koltun 2017; Park et al. 2019; Qi et al. 2018; Wang et al. 2018]. Wang et al. [2018]'s Pix2pixHD model improves both the resolution and the photorealism of Pix2pix [Isola et al. 2017] by employing multi-scale generator and discriminator networks. Recently, Park et al. [2019] propose a spatially-adaptive normalization scheme to better preserve semantic information. Qi et al. [2018] utilize a semi-parametric approach and increase the photorealism of the output images by composing real object segments from a set of training images within an image-to-image synthesis network. Chen and Koltun [2017] try to achieve realism through a carefully crafted regression objective that maps a single input layout to multiple potential scene outputs. Nonetheless, despite modeling one-to-many relationships, the number of outputs is pre-defined and fixed, which still puts tight constraints on the generation process. As compared to these works, besides taking semantic layout as input, our proposed scene generation network is additionally aware of the transient attributes and the latent random noises characterizing intrinsic properties of the generated outputs. As a consequence, our model is more flexible in generating the same scene content under different conditions such as lighting, weather, and seasons.

From training point of view, a careful selection of "negative" pairs, i.e. negative mining, is an essential component in metric learning and ranking [Fu et al. 2013; Li et al. 2013; Shrivastava et al. 2016]. Existing works in CGAN have been using randomly sampled negative image-condition pairs [Reed et al. 2016a]. However, such random negative mining strategy has been shown to be inferior to more meticulous negative sampling schemes [Bucher et al. 2016]. Particularly, the negative pair sampling scheme proposed in our work is inspired by the concept of relevant negative [Li et al. 2013], where the negative examples that are visually similar to positive ones are emphasized more during learning.

To make the generated images look more similar to the reference images, a common technique is to consider feature matching which is commonly employed through a perceptual loss [Chen and Koltun 2017; Dosovitskiy and Brox 2016; Johnson et al. 2016]. The perceptual loss in our proposed model distinguishes itself from existing works by matching segmentation invariant features from pre-trained segmentation networks [Zhou et al. 2017], leading to diverse generations that comply with the given layouts.

2.2 Image Editing

There has been a great effort towards building methods for manipulating visual appearance of a given image. Example-based approaches [Pitie et al. 2005; Reinhard et al. 2001] use a reference image to transfer color space statistics to input image so that visual appearance of input image looks like the reference image. In contrast to these global color transfer approaches, which require highly consistent reference images with input image, user controllable color transfer techniques were also proposed [An and Pellacini 2010; Dale et al. 2009] to consider spatial layouts of input and reference images. Dale et al. [2009] search for some reference images which have

similar visual context to input image in a large image dataset to transfer local color from them and then use color transferred image to restore input image. Other local color transfer approaches [Wu et al. 2013] use the semantic segments to transfer color between regions in reference and input images have same semantic label (e.g. color is transferred from sky region in reference image to sky region in input image). Some data-driven approaches [Laffont et al. 2014; Shih et al. 2013] leverage the time-lapse video datasets taken for same scene to capture scene variations that occur at different times. Shih et al. [2013] aim to give times of day appearances to a given input image, for example converting an input image taken midday to a nice sunset image. They first retrieve the most similar video frame to input scene from dataset as reference frame. Then they find matching patches between reference frame and input image. Lastly, they transfer the variation that occurs between reference frame and desired reference frame which is same scene but taken different time of day to input image. Laffont et al. [2014] take a step forward in their work for handling more general variations as transient attributes such as lighting, weather, and seasons.

High-level image editing offers easier and more natural way to casual users to manipulate a given image. Instead of using a reference image either provided by the user or retrieved from a database, learning the image manipulations and high-level attributes for image editing like a human has also attracted researchers. Berthouzoz et al. [2011] learn parameters of the basic operations for some manipulations recorded in photoshop as macro to adapt them to new images, for example, applying same skin color correction operation with same parameters for both faces with dark-skinned and light-skinned does not give expected correction. In contrast to learning image operations for specific editing effects, Cheng et al. [2014] learn the attributes as adjectives and objects as nouns for semantic parsing of an image and further use them for verbal guided image manipulation to indoor images. For example, the verbal command "*change the floor to wooden*" modifies the appearance of the floor. Similarly, Laffont et al. [2014] learn to recognize transient attributes for attribute-guided image editing on outdoor images. To modify the look of an input image (e.g. a photo taken in a sunny day), they first locate similar scenes in a dataset they collected and annotated with transient attributes. Then they transfer the desired look (e.g. "*more winter*") from the corresponding version of the candidate match images by using an appearance transfer method. Lee et al. [2016] aim to automatically select a subset of style exemplars that will achieve good stylization results by learning a content-to-style mapping between large photo collection and a small style dataset.

Deep learning has fueled a growing literature on employing neural approaches to improve existing image editing problems. Here, we review the studies that are the most relevant to our work. Gatys et al. [2016] have demonstrated how Convolutional Neural Networks (CNNs) effectively encode content and texture separately in feature maps of CNNs trained on large-scale image datasets and have proposed a neural style transfer method to transfer artistic styles from paintings to natural images. Alternatively, Johnson et al. [2016] train a transformation network to speed up the test time of style transferring together with minimization of perceptual loss between input image and stylized image. Li et al. [2017b] consider a deep feed-forward network, which is capable of generating multiple and diverse results within a single network. Recent deep photo style transfer method of Luan et al. [2017], named DPST, aims at providing realism in case of style transfer is made between the real photos. For example, when one wants to make an input photo look like taken in different illumination and weather conditions, a photo-realistic transfer is necessary. It uses semantic labels to prevent semantic inconsistency so that style transfer is carried out between same semantic regions. Recently, Li et al. [2018] have proposed another photo style transfer method called FPST, which works significantly faster than DPST. It considers a two-steps process, a stylization step followed by a photorealistic smoothing step, both of each having efficient closed-form solutions. There are some style transfer networks which are specialized for the editing face images and portraits [Kemelmacher-Shlizerman 2016; Liao et al. 2017; Selim et al. 2016] with new objectives. Nevertheless, these style transfer works limit the users to find an reference photo in which desired style effects exist for desired attributes.

Yan et al. [2016] introduce the first automatic photo adjustment framework based on deep neural networks. They use deep neural network to learn a regressor which transforms the colors for artistic styles especially color adjustment from the image and its stylized version pairs. They define a set of feature descriptors based on pixel, global and semantic levels. In another work, Gharbi et al. [2017] propose a new neural network architecture to learn image enhancement transformations at low resolution, then they move learned transformations to higher resolution in bilateral space in an edge-preserving manner.

Lastly, building upon conditional GAN model, some image completion works have been proposed to predict missing regions providing global and local context information with multiple discriminator networks [Iizuka et al. 2017; Li et al. 2017c].

3 ALS18K DATASET

To train our model, we curate a new dataset by selecting and annotating images from two popular scene datasets, namely ADE20K [Zhou et al. 2017] and Transient Attributes [Laffont et al. 2014], for the reasons which will become clear shortly.

ADE20K [Zhou et al. 2017] includes 22, 210 images from a diverse set of indoor and outdoor scenes which are densely annotated with object and stuff instances from 150 classes. However, it does not include any information about transient attributes. Transient Attributes [Laffont et al. 2014] contains 8, 571 outdoor scene images captured by 101 webcams in which the images of the same scene can exhibit high variance in appearance due to variations in atmospheric conditions caused by weather, time of day, season. The images in this dataset are annotated with 40 transient scene attributes, e.g. sunrise/sunset, cloudy, foggy, autumn, winter, but this time it lacks semantic layout labels.

To establish a richly annotated, large-scale dataset of outdoor images with both transient attribute and layout labels, we further operate on these two datasets as follows. First, from ADE20K, we manually pick the 9,201 images corresponding to outdoor scenes, which contain nature and urban scenery pictures. For these images, we need to obtain transient attribute annotations. To do so,

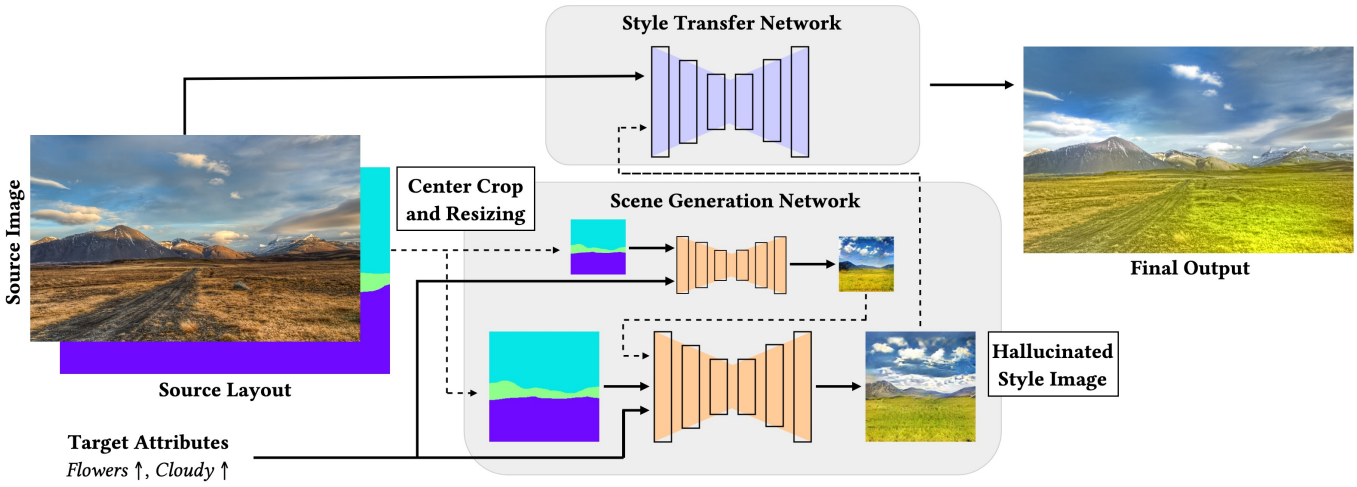


Fig. 2. Overview of the proposed attribute manipulation framework. Given an input image and its semantic layout, we first resize and center-crop the layout to 512 × 512 pixels and feed it to our scene generation network. After obtaining the scene synthesized according to the target transient attributes, we transfer the look of the hallucinated style back to the original input image.

we conduct initial attribute predictions using the pretrained model from [Baltenberger et al. 2016] and then manually verify the predictions. From Transient Attributes, we select all the 8,571 images. To get the layouts, we first run the semantic segmentation model by Zhao et al. [2017], the winner of the MIT Scene Parsing Challenge 2016, and assuming that each webcam image of the same scene has the same semantic layout, we manually select the best semantic layout prediction for each scene and use those predictions as the ground truth layout for the related images.

In total, we collect 17,772 outdoor images (9,201 from ADE20K + 8,571 from Transient Attributes), with 150 semantic categories and 40 transient attributes. Following the train-val split from ADE20K, 8,363 out of the 9,201 images are assigned to the training set, the other 838 testing; for the Transient Attributes dataset, 500 randomly selected images are held out for testing. In total, we have 16,434 training examples and 1,338 testing images. More samples of our annotations are presented in the Supplementary Material. Lastly, we resize the height of all images to 512 pixels and apply center-cropping to obtain 512 × 512 images.

4 ATTRIBUTE MANIPULATION FRAMEWORK

Our framework provides an easy and high-level editing system to manipulate transient attributes of outdoor scenes (see Fig. 2). The key component of our framework is a scene generation network that is conditioned on semantic layout and continuous-valued vector of transient attributes. This network allows us to generate synthetic scenes consistent with the semantic layout of the input image and having the desired transient attributes. One can play with 40 different transient attributes by increasing or decreasing values of certain dimensions. Note that, at this stage, the semantic layout of the input image should also be fed to the network, which can be easily automated by a scene parsing model. Once an artificial scene with desired properties is generated, we then transfer the look of the hallucinated image to the original input image to achieve attribute manipulation in a photorealistic manner.

In Section 4.1, we present the architectural details of our attribute and layout conditioned scene generation network and the methodologies for effectively training our network. Finally, in Section 4.2, we discuss the photo style transfer method that we utilize to transfer the appearance of generated images to the input image.

4.1 Scene Generation

In this section, we first give a brief technical summary of GANs and conditional GANs (CGANs), which provides the foundation for our scene generation network (SGN). We then present architectural details of our SGN model, followed by the two strategies applied for improving the training process. All the implementation details are included in the Supplementary Material.

4.1.1 Background. In Generative Adversarial Networks (GANs) [Goodfellow et al. 2014], a discriminator network D and a generator network G play a two-player min-max game where D learns to determine if an image is real or fake and G strives to output as realistic images as possible to fool the discriminator. The G and D are trained jointly by performing alternating updates:

$$\min_G \max_D \mathcal{L}_{GAN}(G, D) = E_{x \sim p_{data}(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

where x is a natural image drawn from the true data distribution $p_{data}(x)$ and z is a random noise vector sampled from a multivariate Gaussian distribution. The optimal solution to this min-max game is when the distribution p_G converges to p_{data}.

Conditional GANs [Mirza and Osindero 2014] (CGANs) engage additional forms of side information as generation constraints, e.g. class labels [Mirza and Osindero 2014], image captions [Reed et al. 2016b], bounding boxes and object keypoints [Reed et al. 2016a]. Given a context vector c as side information, the generator $G(z, c)$,

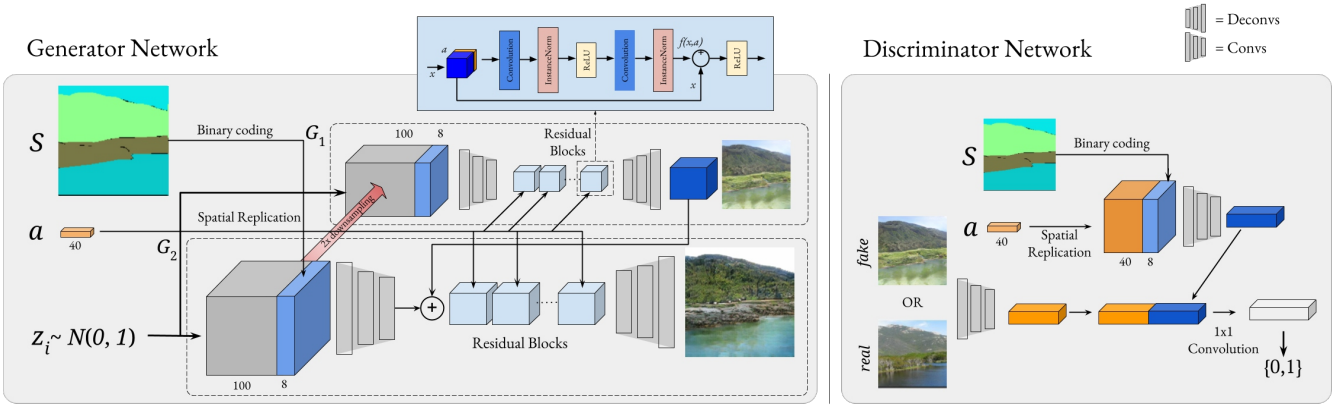


Fig. 3. Our proposed Scene Generation Network (SGN) can generate synthetic outdoor scenes consistent with given layout and transient attributes.

taking both the random noise and the side information, tries to synthesize a realistic image that satisfies the condition c. The discriminator, now having real/fake images and context vectors as inputs, aims at not only distinguishing real and fake images but also whether an image satisfies the paired condition c. Such characteristics are referred to as match-aware [Reed et al. 2016b]. In this way, we expect the generated output of CGAN x_g is controlled by the side information c. Particularly, in our model, c is composed of semantic layouts s and transient attributes a.

4.1.2 Proposed Architecture. In our work, we follow a multi-scale strategy similar to that in Pix2pixHD [Wang et al. 2018]. Our scene generator network (SGN), however, takes the transient scene attributes and a noise vector as extra inputs in addition to the semantic layout. While the noise vector provides stochasticity and controls diversity in the generated images, transient attributes let the users have control on the generation process. In more detail, our multi-scale generator network $G = \{G_1, G_2\}$ consists of a coarse-scale (G_1) generator and a fine-scale (G_2) generator. As illustrated in Fig. 3, G_1 and G_2 have nearly the same architecture, with the exception that they work on different image resolutions. While G_1 operates at a resolution of 256×256 pixels, G_2 outputs an image with a resolution that is $4\times$ larger, i.e. 512×512 pixels. Here, the image generated by G_1 is fed to to G_2 as an additional input in the form of a tensor. In that regard, G_2 can be interpreted as a network that performs local enhancements in the fine resolution.

In our coarse and fine generator networks, while the semantic layout categories are encoded into 8-bit binary codes, transient attributes are represented by a 40-d vector. Input semantic layout map S is of the same resolution with our fine scale image resolution. We concatenate semantic layout S and noise z, and feed their concatenation into convolutional layers of G_1 and G_2 to obtain semantic feature tensors, which are used as input to the subsequent residual blocks. For the coarse scale generator G_1, we at first perform a downsampling operation with a factor of 2 to align the resolutions. Then, spatially replicated attribute vectors a are concatenated to input tensors of each residual block in G_1 and G_2 to condition the image generation process in regard to input transient scene attributes. Finally, deconvolutional layers are used to upsample the feature tensor of the last residual block to obtain final output images. For fine scale generator G_2, semantic feature tensor extracted with the convolutional layers is summed with the feature tensor from the last residual block of coarse generator G_1 before feeding into residual blocks of fine scale generator G_2.

The discriminator used in our SGN also adopts a multi-scale approach in that it includes three different discriminators denoted by D_1, D_2, D_3 with similar network structures that operate at different image scales. In particular, we create an image pyramid of 3 scales that include real and generated high resolution images, their downsampled versions by a factor of 2 and 4. Our discriminators take tuples of real or synthesized images from different levels of this image pyramid, matching or mismatching semantic layouts and transient attributes and decide whether the images are fake or real, and whether the pairings are valid. That is, the discriminator aims to satisfy

$$D_k(x_k, a, S) = \begin{cases} 1, x_k \in p_{\text{data}} \text{ and } x_k, a, S \text{ correctly match,} \\ 0, \text{ otherwise.} \end{cases}$$

with $k = \{1, 2, 3\}$ denoting image scales. Hence, the training our conditional GAN models becomes a multi-task learning problem defined as follows:

$$\min_{G} \max_{D_1, D_2, D_3} \sum_{k=\{1,2,3\}} \mathcal{L}_{GAN}(G, D_k) \tag{2}$$

The architectural details of our Scene Generation Network are given in Table 1. In this table, we follow a naming convention similar to the one used in [Wang et al. 2018; Zhu et al. 2017a]. For instance, $\texttt{C}_3\texttt{128S}_2$ denotes a Convolution-InstanceNorm-ReLU layer with 128 filters of kernel size 3×3 kernel and stride 2. $\texttt{f}_{1\texttt{i}}$ and $\texttt{f}_{2\texttt{i}}$ represent ith internal feature tensors of G_1 and G_2, respectively. $\texttt{R512}$ denotes a residual block with filter size 512 as depicted in Fig. 3. Similarly, $\texttt{D}_3\texttt{128S}_{0.5}$ represents a Deconvolution-InstanceNorm-ReLU layer with 128 filters of kernel size 3×3 and stride 0.5. At the last deconvolution layer $\texttt{D}_7\texttt{3S}_1$, we do not use InstanceNorm and replace ReLU activations with $tanh$. The discriminator resembles a Siamese network [Bromley et al. 1994; Chopra et al. 2005], where one stream

Table 1. Architectural details of the generator and discriminator networks.

Generator		
Net.	Input	Specification
G_1	$z\|\|S$	$\mathtt{C}_7\mathtt{64S}_1 - \mathtt{C}_3\mathtt{128S}_2 - \mathtt{C}_3\mathtt{256S}_2 - \mathtt{C}_3\mathtt{512S}_2 \rightarrow \mathtt{f}_{11}$
	$\mathtt{f}_{11}\|\|a$	$\mathtt{R512} - \mathtt{R512} - \mathtt{R512} - \mathtt{R512} - \mathtt{R512} \rightarrow \mathtt{f}_{12}$
	$\mathtt{f}_{12}$	$\mathtt{D}_3\mathtt{256S}_{0.5} - \mathtt{D}_3\mathtt{128S}_{0.5} - \mathtt{D}_3\mathtt{64S}_{0.5} \rightarrow \mathtt{f}_{13}$
	$\mathtt{f}_{13}$	$\mathtt{D}_7\mathtt{3S}_1 \rightarrow x^{256}_{fake}$
G_2	$z\|\|S$	$\mathtt{C}_7\mathtt{32S}_1 - \mathtt{C}_3\mathtt{64S}_2 \rightarrow \mathtt{f}_{21}$
	$\mathtt{f}_{13} + \mathtt{f}_{21}$	$\mathtt{R64} - \mathtt{R64} \rightarrow \mathtt{f}_{22}$
	$\mathtt{f}_{22}$	$\mathtt{D}_3\mathtt{64S}_{0.5} - \mathtt{D}_7\mathtt{3S}_1 \rightarrow x^{512}_{fake}$
Discriminator		
Net.	Input	Specification
D_k	x	$\mathtt{C}_4\mathtt{64S}_2 - \mathtt{C}_4\mathtt{128S}_2 - \mathtt{C}_4\mathtt{256S}_2 - \mathtt{C}_4\mathtt{512S}_2 \rightarrow \mathtt{f}_x$
	$a\|\|S$	$\mathtt{C}_4\mathtt{64S}_2 - \mathtt{C}_4\mathtt{128S}_2 - \mathtt{C}_4\mathtt{256S}_2 - \mathtt{C}_4\mathtt{512S}_2 \rightarrow \mathtt{f}_c$
	$\mathtt{f}_x\|\|\mathtt{f}_c$	$\mathtt{C}_1\mathtt{512S}_1 - \mathtt{C}_4\mathtt{1S}_1 \rightarrow [0, 1]$

takes the real/generated image as input x and the second one processes the given attributes a and the spatial layout labels S . The responses of these networks are then concatenated $a\|\|S$ and fused via a 1×1 convolution operation. The combined features are finally sent to fully-connected layers for the binary decision. We use leaky ReLU with slope 0.2 for our discriminator networks. We do not use InstanceNorm at the input layers. We employ 3 discriminators at 3 different spatial scales with 1, 0.5 and 0.25 as the scaling factors for both coarse and fine scale generators G_1 and G_2 during training.

4.1.3 Improved Training of SGNs. Here we elaborate on two complementary training techniques that substantially boost the efficiency of the training process.

Relevant Negative Mining. Training the match-aware discriminator in CGAN resembles learning to rank [Rudin and Schapire 2009], in the sense that a "real pair"–real image paired with right conditions–should score higher (i.e. classifying into category 1 in this case) than a "fake pair"–either image is fake or context information is mismatched (i.e. classifying into category 0). For ranking loss, it has been long acknowledged that naively sampling random negative examples is inferior to more carefully designed negative sampling scheme, such as various versions of hard negative mining [Bucher et al. 2016; Fu et al. 2013; Li et al. 2013; Shrivastava et al. 2016]. Analogously, a better negative mining scheme can be employed by training CGAN, as existing works have been using random sampling [Reed et al. 2016a]. To this end, we propose to apply the concept of relevant negative mining [Li et al. 2013] (RNM) to sample mismatching layout in training our SGN model. Concretely, for each layout S, we search for its nearest neighbor S' and set it as the corresponding mismatching negative example for S. In Section 5, we present empirical qualitative and quantitative results to demonstrate improvement from RNM over random sampling. We attempted similar augmentation on attributes a by flipping a few of them instead of complete random sampling to obtain the mismatching a' but found such operation hurt the performance, likely due to the flipped attributes being too semantically close to the original ones which cause ambiguity to the discriminator.

Layout-Invariant Perceptual Loss. Following the practice of existing works [Chen and Koltun 2017; Dosovitskiy and Brox 2016], we also seek to stabilize adversarial training and enhance generation quality by adding a perceptual loss. Conventionally, features used for perceptual loss come from a deep CNN, such as VGG [Simonyan and Zisserman 2014], pretrained on ImageNet for classification task. However, perceptual loss to match such features would intuitively withhold generation diversity, which opposes our intention of creating stochastic output via a GAN framework. Instead, we propose to employ intermediate features trained on outdoor scene parsing with ADE20K. The reason for doing so is three-fold: diversity in generation is not suppressed, because scenes with different contents but the same layout ideally produce the same high-level features; the layout of the generation is further enforced thanks to the nature of the scene parsing network; since the scene parsing network is trained on real images, the perceptual loss will impose additional regularization to make the output more photorealistic. The final version of our proposed perceptual loss is as follows:

$$\mathcal{L}_{percep}(G) = E_{z\sim p_z(z);x,S,a\sim p_{data}(S,a)}\left[\|f_P(x) - f_P(G(z,a,S))\|_2^2\right], \tag{3}$$

where f_P is the CNN encoder for the scene parser network. Our full objective that combines multi-scale GAN loss and layout-invariant feature matching loss thus becomes:

$$\min_G\left(\left(\max_{D=\{D_1,D_2,D_3\}}\sum_{k=1,2,3}\mathcal{L}_{GAN}(G,D_k)\right) + \lambda\mathcal{L}_{percep}(G)\right) \tag{4}$$

where λ is a scalar controlling the importance of our proposed layout-invariant feature matching loss and is set to 10 in our experiments. By additionally considering RNM and perceptual loss, we arrive at the training procedure which is outlined in Algorithm 1.

Algorithm 1: SGN training algorithm

Input: Training set $\Omega = \{(x, a, S)\}$ with training images x, semantic segmentation layouts S and transient attributes a.
for all number of iterations **do**
  sample minibatch of paired x, a, S
  sample minibatch of z_i from $\mathcal{N}(0, I)^Z$
  for all (x_i, a_i, S_i) in Ω **do**
    Randomly sample negative a'_i mismatching x_i
    Sample S'_i mismatching x_i via RNM
  end for
  $x_g \leftarrow G(z_i, a_i, S_i)$ {Forward through generator}
  for k=1:3 **do**
    $\mathcal{L}_{D_k} \leftarrow -(\log D_k(x, a, S) + \log(1 - D_k(x_g, a, S)) + \log(1 - D_k(x, a', S'))$
    $D_k \leftarrow D_k - \alpha\partial\mathcal{L}_{D_k}/\partial D_k$ {Update discriminator D_k}
  end for
  $\mathcal{L}_G \leftarrow -\log D(x_g, a, S) + \lambda\|f_p(x) - f_p(x_g)\|_2^2$
  $G \leftarrow G - \alpha\partial\mathcal{L}_G/\partial G$ {Update generator G}
end for

4.2 Style Transfer

The main goal in photo style transfer is to successfully transfer visual style (such as color and texture) of a reference image onto another image while preserving semantic structure of the target image. In the past, statistical color transfer methods [Pitie et al. 2005; Reinhard et al. 2001] showed that the success of the style transfer methods highly depend on the semantic similarity of the source and target images. To overcome this obstacle, user interaction, semantic segmentation approaches or image matching methods were utilized to provide semantic relation between source and target images. In addition, researchers explored data driven methods to come up with fully automatic approaches which retrieve the source style image through some additional information such as attributes, features and semantic similarity.

For existing deep learning based photo style transfer methods, it is still crucial that source and reference images have similar semantic layouts to provide successful and realistic style transfer results. Image retrieval based approaches are limited with the dataset and they become infeasible when there is no images with the desired properties. The key distinguishing characteristics of our framework is that we can generate a style image on the fly that has both similar semantic layout with the input image and possess the desired transient attributes, thanks to our proposed SGN model. In our framework, for photo style transfer, we consider employing both DPST [Luan et al. 2017] and FPST [Li et al. 2018] models.

DPST [Luan et al. 2017] extends the formalization of the neural style transfer method of Gatys et al. [2016] by adding a photorealism regularization term that enables the style transfer to be done between same semantic regions instead of the whole image. This property makes DPST very appropriate for our image manipulation system. Although this method in general produces fairly good results, we observe that it sometimes introduces some smoothing and visual artifacts in the output images, which hurt the photorealism. For that reason, we first apply a cross bilateral filter [Chen et al. 2007] to smooth the DPST's output according to edges in the input image and then apply the post-processing method proposed by Mechrez et al. [2017], which uses screened Poisson equation to make the stylized image more similar to the input image in order to increase its visual quality.

FPST [Li et al. 2018] formulates photo style transfer as a two steps procedure. The first step carries out photorealistic image stylization by using a novel network architecture motivated by the whitening and coloring transform [Li et al. 2017a], in which the upsampling layers are replaced with unpooling layers. The second step performs a manifold ranking based smoothing operation to eliminate the structural artifacts introduced by the first step. As both of these steps have closed-form solutions, FPST works much faster than DPST. Since FPST involves an inherent smoothing step, in our experiments, we only apply the approach by Mechrez et al. [2017] as a post-processing step.

5 RESULTS AND COMPARISON

We first evaluate our scene generation network's ability to synthesize diverse and realistic-looking outdoor scenes, then show attribute manipulation results of our proposed two-stage framework that employs the hallucinated scenes as reference style images. Lastly, we discuss the limitations of the approach.

5.1 Attribute and Layout Guided Scene Generation

Here, we assess the effectiveness of our SGN model on generating outdoor scenes in terms of image quality, condition correctness and diversity. We also demonstrate how the proposed model enables the users to add and subtract scene elements.

5.1.1 Training Details. All models were trained with a mini-batch size of 40 where parameters were initialized from a zero-centered Gaussian distribution with standard deviation of 0.02. We set the amount of the layout-invariant feature matching loss λ to 10. We used the Adam optimizer [Kingma and Ba 2014] with the learning rate value of 2×10^{-4} and the momentum value of 0.5. For data augmentation, we employed horizontal flipping with a probability of 0.5. We trained our coarse-scale networks for 100 epochs on a NVIDIA Tesla K80 GPU for 3 days. After training them, we kept their parameters fixed and trained our fine-scale networks for 10 epochs. Then, in the next 70 epochs, we updated the parameters of both of our fine and coarse-scale networks together. Our implementation is based on the PyTorch framework. Training of our fine-scale networks took about 10 days on a single GPU.

5.1.2 Ablation Study. We illustrate the role of Relevant Negative Mining (RNM) and layout-invariant Perceptual Loss (PL) in improving generation quality with an ablation study. Here we consider the outputs of the coarse-scale generator G_1 to evaluate these improvements as it acts like a global image generator. Our input layouts come from the test set, i.e. are unseen during training. Furthermore, we fix the transient attributes to the predictions of the pre-trained deep transient model [Baltenberger et al. 2016]. Fig. 4 presents synthetic outdoor images generated from layouts depicting different scene categories such as urban, mountain, forest, coast, lake and highway. We make the following observations from these results.

Attributes of the generated images are mostly in agreement with the original transient attributes. Integrating RNM slightly improves the rendering of attributes but in fact, its main role is to make training more stable. Our proposed layout-invariant PL boosts the final image quality of SGN. The roads, the trees and the clouds are drawn with the right texture; the color distributions of the sky, the water and the field also appear realistic; reasonable physical effects are also observed such as the reflection of the water, fading of the horizon, valid view perspective of urban objects. In our analysis, we also experimented with the VGG-based perceptual loss, commonly employed in many generative models, but as can be seen from Fig. 4, our proposed perceptual loss, which performs feature matching over a pretrained segmentation network, gives much better results in terms of photorealism. Overall, the results with both RNM and PL are visually more pleasing and faithful to the attributes and layouts.

For quantitative evaluation, we employ the Inception Score (IS) [Salimans et al. 2016] and the Fréchet Inception Distance (FID) [Heusel et al. 2017][2]

[2]In our evaluation, we utilized the official implementations of IS and FID. IS scores are estimated by considering all of the test images from our dataset, which were not seen

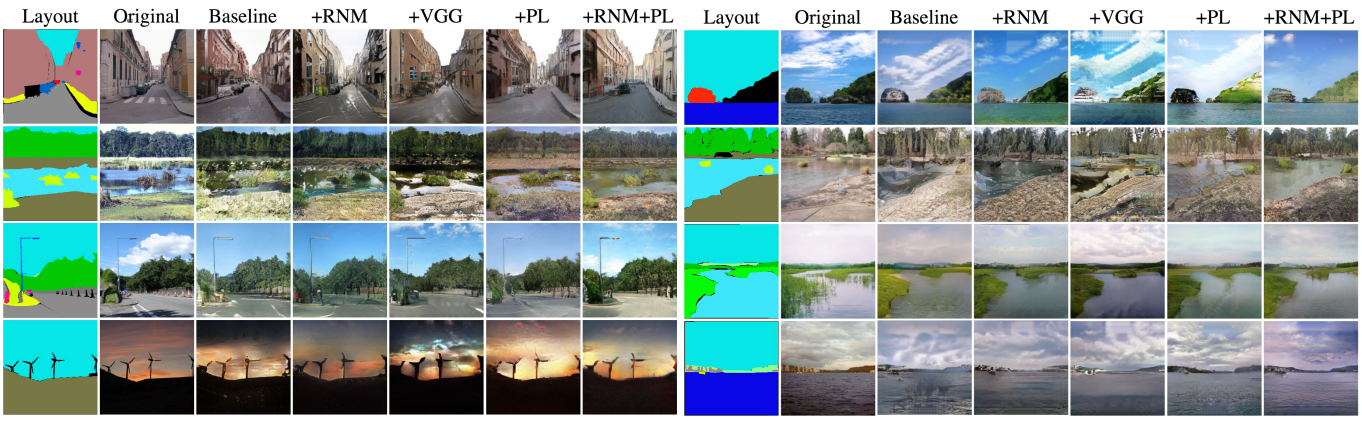


Fig. 4. Sample scene generation results. In these examples, the input layouts are from the test set, which are unseen during training and the transient attributes are fixed to the original transient attributes. Incorporating Relevant Negative Mining (RNM) and Perceptual Loss (PL) significantly improves the performance of the baseline SGN model in terms of both image quality as well as faithfulness of the end result to conditioned layouts and attributes. Moreover, the way we define our perceptual loss, as compared to commonly used VGG-based one, provides better and more photorealistic results.

The IS correlates well with human judgment of image quality where higher IS indicates better quality. FID has been demonstrated to be more reliable than IS in terms of assessing the realism and variation of the generated samples. Lower FID value means that the distributions of generated images and real images are similar to each other. Table 2 shows the IS and FID values for our SGN model trained under various settings, together with values for the real image space. These results agree with our qualitative analysis that training with RNM and Perceptual Loss provides samples of the highest quality. Additionally, for each generated image, we also predict its attributes and semantic segmentation map using separately trained attribute predictor by Baltenberger et al. [2016] and the semantic segmentation model by Zhou et al. [2017] and we report the average MSE[3] and segmentation accuracy again in Table 2. Training with the proposed perceptual loss is more effective in reflecting photorealism and preserving both the desired attributes and the semantic layout better than the VGG-based perceptual loss.

Our SGN model with RNM and Perceptual Loss shows clear superiority to other variants both qualitatively and quantitatively. Thus from now on, if not mentioned otherwise, all of our results are obtained with this model.

5.1.3 Comparison with Image-to-Image Translation Models. We compare our model to Pix2pix [Isola et al. 2017] and Pix2pixHD [Wang et al. 2018] models[4]. It is worth mentioning that both of these two approaches generate images only by conditioning on the semantic layout but not transient attributes, and moreover, they do not utilize noise vectors. We provide qualitative comparisons in Fig. 5. As these results demonstrate, our model not only generates realistic looking images on par with Pix2pixHD but also has the capability to deliver control over the attributes of the generated scenes. "Sunset" attribute makes the horizon slightly more reddish, "Dry" attribute increases the brown tones on the trees, "Snow" attribute whitens the ground. Also note that the emergence of each attribute tends to highly resonate with part of the image that is most related to the attribute. That is, "Clouds" attribute primarily influences the sky, whereas "Winter" attribute correlates with the ground, and "Lush" tends to impact the trees and the grass. This further highlights our model's reasoning capability about the attributes in producing realistic synthetic scenes.

Table 2. Ablation study. We compare visual quality with respect to Inception Score (IS) and Fréchet Inception distance (FID), attribute and semantic layout correctness in terms of average MSE of attribute predictions (Att. MSE) and segmentation accuracy (Seg. Acc.), respectively, via pre-trained models. Our SGN model trained with RNM and PL techniques consistently outperforms the others, including the setting with VGG-based perceptual loss.

Model	IS	FID	Att. MSE	Seg. Acc.
SGN	3.91	43.77	0.016	67.70
+RNM	3.89	41.84	0.016	70.11
+VGG	3.80	41.87	0.016	67.42
+PL	4.15	36.42	**0.015**	70.44
+RNM+PL	**4.19**	**35.02**	**0.015**	**71.80**
Original	5.77	0.00	0.010	75.64

For quantitative comparison, we compare the IS and FID scores and segmentation accuracy using all 1,338 testing images in Table 3 considering both coarse and fine scales. These results suggest that our proposed model produces high fidelity natural images better than Pix2pixHD in both scales. The difference in the segmentation accuracy suggests that Pix2pixHD puts a more strict restraint on the layout whereas our model offers flexibility in achieving a reasonable trade-off between capturing realism in accordance with transient attributes vs. fully agreeing with the layout. Furthermore, in addition to these metrics, we conduct a human evaluation on Figure Eight[5],

during training and by using a split size of 10. While calculating FID scores, we employ all of the test images from our dataset as the reference images.

[3]The ground truth attributes are scalar values between 0 and 1.

[4]For both of these models, we use the original source codes provided by the authors.

[5]Figure Eight is a web-based data annotation company which can be accessed from https://www.figure-eight.com/

Fig. 5. Comparison of our SGN model against Pix2pix [Isola et al. 2017] and Pix2pixHD [Wang et al. 2018]. Each row shows the original image and the samples generated according to its corresponding semantic layout. Since our SGN model also takes into account a set of target transient attributes (only the top three most significant ones are shown here for the sake of simplicity), it can generate diverse and more realistic results than the other methods.

Table 3. Quantitative comparison of layout conditioned image synthesis approaches. Our model consistently outperforms others in both coarse and fine resolutions in terms of photorealism, as measured by IS and FID.

	Model	IS	FID	Seg. Acc.
Coarse	Pix2pix	3.26	76.40	61.93
	Pix2pixHD	4.20	47.86	**75.57**
	Ours	**4.19**	**35.02**	71.80
	Original	5.77	0.00	75.64
Fine	Pix2pixHD	4.87	50.85	**76.17**
	Ours	**5.05**	**36.34**	74.60
	Original	7.37	0.00	77.14

asking workers to select among the results of our proposed model and the Pix2pixHD method (for the same semantic layout) which they believe is more realistic. We randomly generate 200 questions, and let 5 different subjects answer each question. We provide the details of our user study in the Supplementary Material. We find that 66% of the subjects picked our results as more realistic. These results suggest that besides the advantages of manipulation over transient attributes, our model also produces higher quality images than the Pix2pixHD model. We also compared our results to the recently proposed Cascaded Refinement Network [Chen and Koltun 2017], however, it did not give meaningful results on our dataset with complex scenes[6].

5.1.4 Diversity of the Generated Images. In our framework, a user can control the diversity via three different mechanisms, each playing a different role in the generation process. Perhaps the most important one is the input semantic layout which explicitly specifies the content of the synthesized image, and the other two are the target transient attributes and the noise vector. In Fig. 6, we show the effect of varying the transient attributes for a sample semantic layout and Fig. 7 illustrates the role of noise. If we keep the layout and the attributes fixed, the random noise vector mainly affects the appearance of some local regions, especially the ones involving irregular or stochastic textures such as the sky, the trees or the plain grass. The transient attribute vectors, however, have a more global effect, modifying the image without making any changes to the constituent parts of the scene.

5.1.5 Adding and Subtracting Scene Elements. Here we envision a potential application of our model as a scene editing tool that can add or subtract scene elements. Fig. 8 demonstrates an example. We begin with a coarse spatial layout which contains two large segments denoting the "sky" and the "ground". We then gradually add new elements, namely "mountain", "tree", "water". At each step, our model inserts a new object based on the semantic layout. In fact, such a generation process closely resembles human thought process in imagining and painting novel scenes. The reverse process, subtracting elements piece by piece, can be achieved in a similar manner. We sample different random attribute vectors to illustrate how generation diversity can enrich the outcomes of such photo-editing tools and provide a video demo in the Supplementary Material.

5.2 Attribute Transfer

We demonstrate our attribute manipulation results in Fig. 9. Here we only provide results obtained by using FPST [Li et al. 2018] as it

[6]We trained this model using the official code provided by the authors.

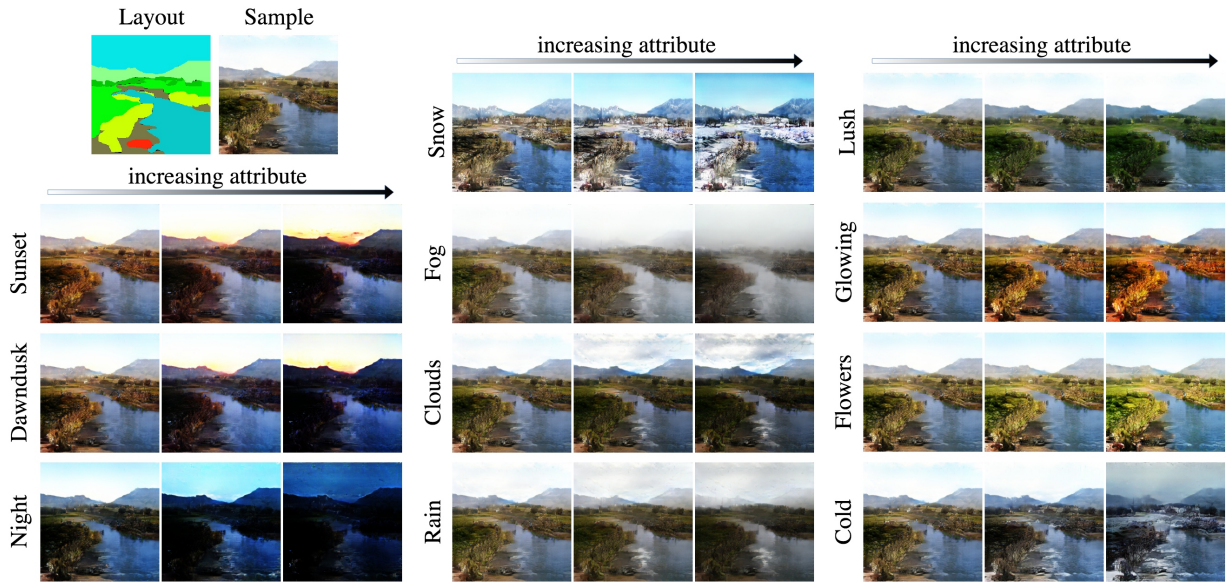
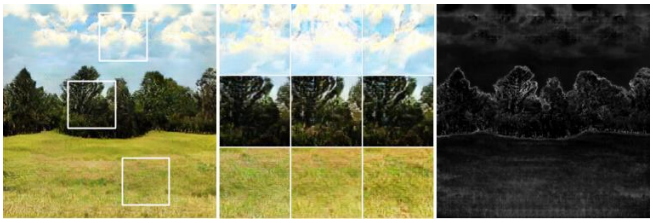


Fig. 6. Modifying transient attributes in generating outdoor images under different weather and time conditions. Our model's ability of varying with transient attributes contributes to the diversity and photorealism in its generation (more results can be found in the Supplementary Material).



A synthetic scene Stochastic variations Standard deviation

Fig. 7. Effect of the noise vector. For an example synthetically generated scene (left), we show close-up views from three different regions (middle) from samples obtained with only changing the random noise. Standard deviation of each pixel over 100 different realizations of the scene (right), which demonstrates that the random noise causes stochastic variations within the irregular or stochastic textural regions.

gives slightly better results in our experiments and also significantly faster than DPST [Luan et al. 2017]. From now on, unless otherwise stated, all of our attribute transfer results will be the ones obtained with FPST. We provide the results of DPST in the Supplementary Material. As can be seen, our algorithm produces photorealistic manipulation results for many different types of attributes like "Sunset", "Spring", "Fog", "Snow", and moreover, a distinctive property of our approach is that it can perform multimodal editing for a combination of transient attributes as well, such as "Winter and Clouds" and "Summer and Moist". It should be noted that modifying an attribute is inherently coupled with the appearance of certain semantic scene elements. For example, increasing "Winter" attribute makes the color of the grass white whereas increasing "Autumn" attribute turns them to brown. As another example, "Clouds" attribute does not modify the global appearance of the scene but merely the sky region, comparing with "Fog" attribute which blurs distant objects; "Dry" attribute emphasizes the hot colors, while "Warm" attribute has the opposite effect. Some attributes such as "Fog", however, have an influence on the global appearance.

In Fig. 10, we compare the performance of our method to the data-driven approach of Laffont et al. [2014]. As mentioned in Section 2, this approach first identifies a scene that is semantically similar to the input image using a database of images with attribute annotations, then it retrieves the version of that scene having the desired properties, and finally, the retrieved image is used as a reference for style transfer. For retrieving the images semantically similar to the source image we also use the Transient Attributes dataset and the retrieval strategy employed by Laffont et al. [2014]. In fact, since the authors did not publicly share their attribute transfer code, in our experiments, we consider the test cases provided in their project website [7]. In the figure, we both present the reference images generated by our approach and retrieved by the competing method at the right-bottom corner of each output image. For a fair comparison, we also present alternative results of [Laffont et al. 2014] where we replace the original exemplar-based transfer method with FPST [Li et al. 2018], which is used in obtaining our results[8]. As can be seen, our approach produces better results than [Laffont et al. 2014] in terms of visual quality and as to reflecting the desired transient attributes. These results also demonstrate how style transfer methods are dependent on semantic similarity between the input and style images. Our main advantage over the approach by Laffont et al. [2014] is that the target image is directly hallucinated from the source image via the proposed SGN model, instead of retrieving

[7] The test cases we used in our experimental analysis are available at http://transattr.cs.brown.edu/comparisonAppearanceTransfer/testCases.html.
[8] Note that, the post-processing method Mechrez et al. [2017] is also employed here to improve photorealism.

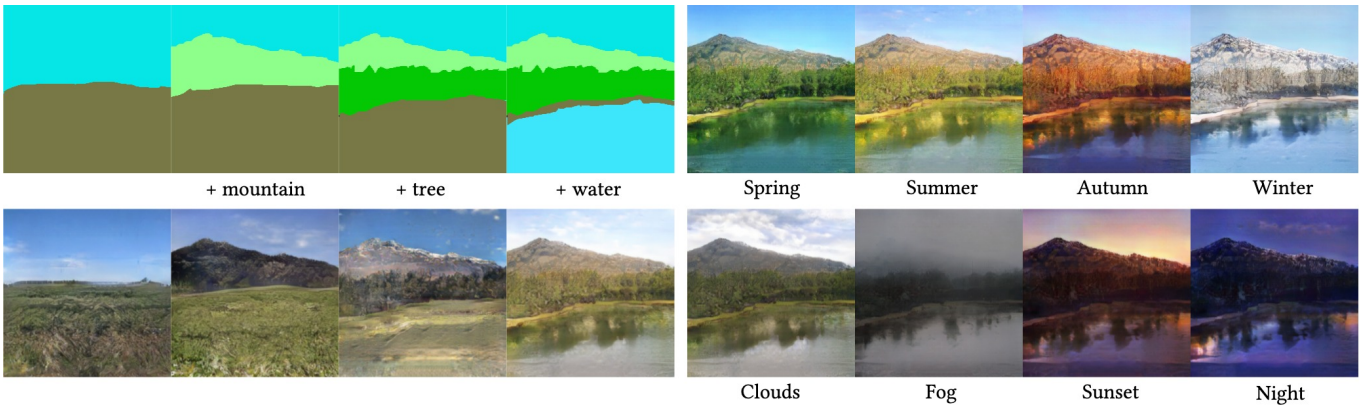


Fig. 8. Gradually adding and removing elements to and from the generated images. We use a coarse spatial layout map (top left) to generate an image from scratch, and then keep adding new scene elements to the map to refine the synthesized images. Moreover, we also show how we can modify the look by conditioning on different transient attributes.

Fig. 9. Sample attribute manipulation results. Given an outdoor scene and its semantic layout, our model produces realistic looking results for modifying various different transient attributes. Moreover, it can perform multimodal editing as well, in which we modify a combination of attributes.

the target image from a training set. This makes a difference since the source and the target images always share the same semantic layout. In this regard, our approach provides a more natural way to edit an input image to modify its look under different conditions.

Additionally, we conducted a user study on Figure Eight to validate our observations. We show the participants an input image and a pair of manipulation results along with a target attribute and force them to select one of the manipulated images which they

Fig. 10. Comparison with [Laffont et al. 2014]. In each row, for a given input image (first column), we respectively provide the results of [Laffont et al. 2014] using their exemplar-based style transfer method (second column) and FPST [Li et al. 2018] (third column) between retrieved images and input images, and the results of our method (last column) using FPST [Li et al. 2018] between generated image by proposed SGN model and input image.

Table 4. User study results for attribute manipulation. The preference rate denotes the percentage of comparisons in which users favor one method over the other.

	Preference rate
Ours w/ FPST > Laffont et al. [2014]	65%
Ours w/ FPST > Laffont et al. [2014] w/ FPST	83%
Ours w/ FPST > Ours w/ DPST	52%

consider visually more appealing regarding the specified target attribute. The manipulation results are either our results obtained by using DPST or FPST, or those of [Laffont et al. 2014]. We have a total of 60 questions and we collected at least 3 user responses per each of these question. We provide the details of our user study in the Supplementary Material. Table 4 summarizes these evaluation results. We find that the human subjects prefer our approach against the data-driven approach by [Laffont et al. 2014] 65% of the time. This margin substantially increases when we replace the original exemplar-based transfer part of [Laffont et al. 2014] with FPST as the semantic layouts of retrieved images are most of the time not consistent with those of the input images. We also evaluate the results of our frameworks with FPST and DPST being used as the style transfer network. As can be seen from Table 4, the human subjects prefer FPST against DPST but by a very small margin.

The most important advantage of our framework over existing works is that our approach enables users to play with the degree of desired attributes via changing the numerical values of the attribute condition vector. As shown in Fig. 11, we can increase and decrease the strength of specific attributes and smoothly walk along the learned attribute manifold using the outputs from the proposed SGN model. This is nearly impossible for a retrieval-based editing system since the style images are limited with the richness of the database.

Although our attribute manipulation approach is designed for natural images, we can apply it to oil paintings as well. In Fig. 12, we manipulate transient attributes of three oil paintings to obtain their novel versions depicting these landscapes at different seasons. As can be seen from these results, our model also gives visually

Fig. 11. Our method can produce photorealistic manipulation results for different degrees of transient attributes.

Fig. 12. Season transfer to paintings. Source images: Wheat Field with Cypresse by Vincent van Gogh (1889), In the Auvergne by Jean-Francois Millet (1869) and Lourmarin by Paul-Camille Guigou (1868), respectively.

pleasing results for these paintings, hallucinating how they might look like if the painters picture the same scene at different times.

5.2.1 Effect of Post-Processing and Running Times. We show the effects of the post-processing steps involved in our framework in Fig. 13. As mentioned in Section 4.2, for DPST based stylized images, we first apply a cross bilateral filter (BLF) [Chen et al. 2007] and then employ screened Poisson equation (SPE) based photorealism enhancement approach [Mechrez et al. 2017]. For FPST based stylized images, we only apply SPE as it inherently performs smoothing. As can be seen from these results, the original stylized images demonstrate some texture artifacts and look more like a painting. Our post-processing steps make these stylized images photorealistic and more similar to the given input image.

In Table 5, we provide the total running time of our framework for manipulating the attributes of an outdoor image. There are three main parts, namely the scene generator network (SGN), the style transfer network, and the post-processing. We report the running time of each of these steps as well. For the style transfer and the post-processing steps, we employ two different versions, one depends on DPST and the other one depends on FPST, and the corresponding smoothing operations. The experiment is conducted on a system with an NVIDIA Tesla K80 graphics card. We consider three different sizes for the input image and report the average run-time for each

Fig. 13. Effect of post-processing. Top: a sample input image and "Autumn" attribute transfer results by our framework with DPST [Luan et al. 2017] and FPST [Li et al. 2018], respectively. Bottom: the impact of various post-processing strategies on the final results. See Section 4.2 for the details.

Table 5. Running time analysis showing the average run time (in seconds) of each component of the proposed model across various image resolutions.

	Resolution	SGN	Style Tranfer	Post-Processing	Total
DPST	512 × 256	0.10	1245.31	2.52	1247.93
	768 × 384	0.10	2619.48	4.61	2626.19
	1024 × 512	0.10	4130.27	7.24	4137.51
FPST	512 × 256	0.10	36.54	1.54	38.18
	768 × 384	0.10	99.34	3.63	103.07
	1024 × 512	0.10	222.20	6.22	228.52

image resolution. Our FPST-based solution is, in general, much faster than our DPST-based one as most of the computation time is spent on the style transfer step. For images of 1024 × 512 pixels, while it takes 4 minutes to manipulate the attributes of an image with FPST, DPST requires 70 minutes to achieve the task.

5.3 Effect of Center-cropping

Our SGN works with a fixed resolution of 512 × 512 pixels and accepts the semantic layout of the center cropped and resized version of the input image. The style transfer networks consider SGN's output as the target style image and manipulates the input image accordingly. When the image is very wide, like a panorama, center-cropping omits most of the source image. We analyze how this affects the overall performance our framework on a couple of panoramic images from SUN360 dataset [Xiao et al. 2012]. We present attribute manipulation results for one of these images in Fig. 14 and present the rest in the Supplementary Material. We have observed that center-cropping does not pose a serious drawback to our approach, since the style transfer step exploits semantic layouts to constrain color transformations to be carried out between features from the image regions with the same label.

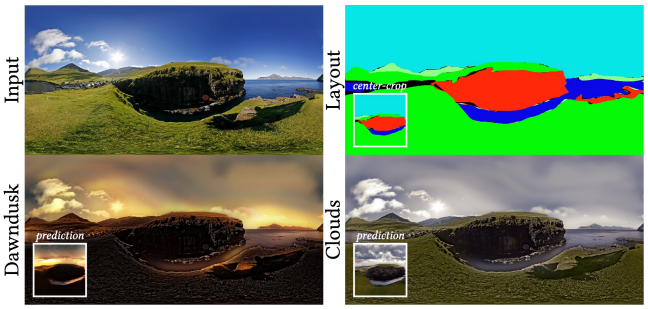


Fig. 14. Effect of center-cropping on manipulating a panoramic image.

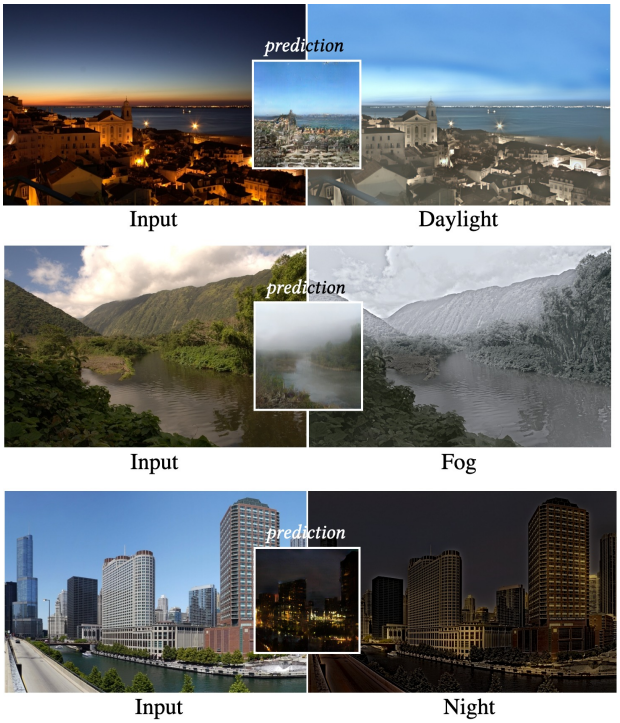


Fig. 15. Example failure cases for our attribute manipulation framework, which are due to the visual quality of synthesized reference style image (top row) and failing of the photo style transfer method (bottom two rows).

5.4 Limitations

Our framework generally gives quite plausible results, but we should note that it might fail in some circumstances if either one of its components fails to function properly. In Fig. 15, we demonstrate such example failure cases. In the first row, the photo-realistic quality of the generated scene is not very high as it does note reproduce the houses well. As a consequence, the manipulation result is not very convincing. For the last two scenes, our SGN model hallucinated "Fog" and "Night" attributes successfully but the style transfer network fails to transfer the looks to the input images.

6 CONCLUSION

We have presented a high-level image manipulation framework to edit transient attributes of natural outdoor scenes. The main novelty of the paper is to utilize a scene generation network in order to synthesize on the fly the reference style image that is consistent with the semantic layout of the input image and exhibit the desired attributes. Trained on our richly annotated ALS18K dataset, the proposed generative network can hallucinate many different attributes reasonably well and even allows edits with multiple attributes in a unified manner. For future work, we plan to extend our model's functionality to perform local edits based on natural text queries, e.g. add or remove certain scene elements using referring expressions. Another interesting and more challenging research direction is to replace the proposed two-staged model with an architecture that can perform the manipulation in a single shot.

ACKNOWLEDGMENTS

This work was supported in part by TUBA GEBIP fellowship awarded to E. Erdem. We would like to thank NVIDIA Corporation for the donation of GPUs used in this research. This work has been partially funded by the DFG-EXC-Nummer 2064/1-Projektnummer 390727645.

REFERENCES

Xiaobo An and Fabio Pellacini. 2010. User-controllable color transfer. In *Computer Graphics Forum*, Vol. 29. Wiley Online Library, 263–271.

G. Antipov, M. Baccouche, and J. Dugelay. 2017. Face aging with conditional generative adversarial networks. In *The IEEE International Conference on Image Processing (ICIP)*.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *arXiv preprint arXiv:1701.07875* (2017).

Ryan Baltenberger, Menghua Zhai, Connor Greenwell, Scott Workman, and Nathan Jacobs. 2016. A fast method for estimating transient scene attributes. In *Winter Conference on Application of Computer Vision (WACV)*.

David Berthelot, Tom Schumm, and Luke Metz. 2017. BEGAN: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* (2017).

Floraine Berthouzoz, Wilmot Li, Mira Dontcheva, and Maneesh Agrawala. 2011. A Framework for content-adaptive photo manipulation macros: Application to face, landscape, and global manipulations. *ACM Transactions on Graphics (TOG)* 30, 5 (2011), 120–1.

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature Verification using a "Siamese" Time Delay Neural Network. In *In Advances in Neural Information Processing Systems (NeurIPS)*.

Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. 2016. Hard negative mining for metric learning based zero-shot classification. In *European Conference on Computer Vision Workshops (ECCVW)*.

Jiawen Chen, Sylvain Paris, and Frédo Durand. 2007. Real-time edge aware image processing with the bilateral grid. *ACM Transactions on Graphics (TOG)* 26, 3 (2007).

Qifeng Chen and Vladlen Koltun. 2017. Photographic image synthesis with cascaded refinement networks. In *The IEEE International Conference on Computer Vision (ICCV)*.

Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J Mitra, and Philip Torr. 2014. ImageSpirit: Verbal guided image parsing. *ACM Transactions on Graphics (TOG)* 34, 1 (2014), 3.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kevin Dale, Micah K Johnson, Kalyan Sunkavalli, Wojciech Matusik, and Hanspeter Pfister. 2009. Image restoration using online photo collections. In *The IEEE International Conference on Computer Vision (ICCV)*.

Alexey Dosovitskiy and Thomas Brox. 2016. Generating images with perceptual similarity metrics based on deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yifan Fu, Xingquan Zhu, and Bin Li. 2013. A survey on instance selection for active learning. *Knowledge and information systems* 35, 2 (2013), 249–283.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *The IEEE Computer Vision and Pattern Recognition (CVPR)*.

Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. 2017. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 118.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. 2015. DRAW: A recurrent neural network for image generation. In *International Conference on Machine Learning (ICML)*.

Ishaan Gulrajani, Kundan Kumar, Faruk Ahmed, Adrien Ali Taiga, Francesco Visin, David Vazquez, and Aaron Courville. 2016. PixelVAE: A latent variable model for natural images. *arXiv preprint arXiv:1611.05013* (2016).

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *The IEEE International Conference on Computer Vision (ICCV)*.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 107.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*.

Ira Kemelmacher-Shlizerman. 2016. Transfiguring portraits. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 94.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*.

Pierre-Yves Laffont, Zhile Ren, Xiaofeng Tao, Chao Qian, and James Hays. 2014. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (TOG)* 33, 4 (2014).

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *International Conference on International Conference on Machine Learning (ICML)*.

Joon-Young Lee, Kalyan Sunkavalli, Zhe Lin, Xiaohui Shen, and In So Kweon. 2016. Automatic content-aware color and tone stylization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xirong Li, CeesG M Snoek, Marcel Worring, Dennis Koelma, and Arnold WM Smeulders. 2013. Bootstrapping visual categorization with relevant negatives. *IEEE Transactions on Multimedia* 15, 4 (2013), 933–945.

Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.H. Yang. 2017a. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017b. Diversified texture synthesis with feed-forward networks. In *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 3920–3928.

Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. 2018. A closed-form solution to photorealistic image stylization. In *European Conference on Computer Vision (ECCV)*. 453–468.

Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. 2017c. Generative face completion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1. 3.

Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. 2017. Visual attribute transfer through deep image analogy. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 120.

Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. 2017. Deep photo style transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Elman Mansimov, Emilio Parisotto, Lei Jimmy Ba, and Ruslan Salakhutdinov. 2016. Generating images from captions with attention. In *International Conference on Learning Representations (ICLR)*.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, and Zhen Wang. 2016. Multi-class generative adversarial networks with the L2 loss function. *arXiv preprint arXiv:1611.04076* (2016).

Roey Mechrez, Eli Shechtman, and Lihi Zelnik-Manor. 2017. Photorealistic style transfer with screened Poisson equation. In *The British Machine Vision Conference (BMVC)*.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional Image Generation with PixelCNN Decoders. *arXiv preprint arXiv:1606.05328* (2016).

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. 2005. N-dimensional probability density function transfer and its application to color transfer. In *The IEEE International Conference on Computer Vision (ICCV)*.

Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. 2018. Semi-parametric image synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. 8808–8816.

Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks.. In *International Conference on Learning Representations (ICLR)*.

Scott Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016a. Learning what and where to draw. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016b. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. 2001. Color transfer between images. *IEEE Computer graphics and applications* 21, 5 (2001), 34–41.

Cynthia Rudin and Robert E Schapire. 2009. Margin-based ranking and an equivalence between AdaBoost and RankBoost. *Journal of Machine Learning Research* 10, Oct (2009), 2193–2232.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2234–2242.

Ahmed Selim, Mohamed Elgharib, and Linda Doyle. 2016. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 129.

Wenling Shang, Kihyuk Sohn, Zeynep Akata, and Yuandong Tian. 2017. Channel-Recurrent Variational Autoencoders. *arXiv preprint arXiv:1706.03729* (2017).

Yichang Shih, Sylvain Paris, Frédo Durand, and William T Freeman. 2013. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Transactions on Graphics (TOG)* 32, 6 (2013), 200.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional GANs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–13.

Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised learning of visual representations using videos. In *The IEEE International Conference on Computer Vision (ICCV)*.

Fuzhang Wu, Weiming Dong, Yan Kong, Xing Mei, Jean-Claude Paul, and Xiaopeng Zhang. 2013. Content-based colour transfer. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 190–203.

J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba. 2012. Recognizing scene viewpoint using panoramic place representation. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2695–2702.

Zhicheng Yan, Hao Zhang, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. 2016. Automatic photo adjustment using deep neural networks. *ACM Transactions on Graphics (TOG)* 35, 2 (2016), 11.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ADE20K dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*. 1558–1566.

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017b. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems (NeurIPS)*. 465–476.